The Thesis Committee for Liang Zhu
Certifies that this is the approved version of the following thesis:

Specialization in the Identity Ecosystem

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Kathleen Suzanne Barber

Sarfraz Khurshid

Specialization in the Identity Ecosystem

by

Liang Zhu, B.S.; M.E.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

December 2014

Dedication

To my parents, Renchun Zhu, and Xiaolan Yang, who are always by my side. To my grandpa, Fufeng Yang, who is a retired elementary school teacher and always encourages me to pursue a graduate degree.

Acknowledgements

I would like to thank my supervisor, Dr. Suzanne Barber, who always guides my research with patience and financially supports me during the past year.

Thanks to Dr. Khurshid, for consenting to be in the supervising committee and providing valuable suggestions.

Thanks to Dr. Muhammad Zubair Malik, his guidance in my research helps me greatly. Thanks to Yongpeng Yang, Shayani Deb, Monisha Manoharan, Mario Guel for their valuable suggestions and feedbacks during problem discussions.

Thanks to Dr. Dina Inman Ramgolam, Dr. Andy Maloney, Dr. Razieh Nokhbeh Zaeem, Tara Upchurch, and all other members in the Center of Identity who are so nice and helpful.

Special thanks to my parents for their endless love and always be my support.

Abstract

Specialization in the Identity Ecosystem

Liang Zhu, M.S.E.

The University of Texas at Austin, 2014

Supervisor: Kathleen Suzanne Barber

Cyberspace has dramatically improved our daily lives in the past several decades. Meanwhile, people's personal identifiable information (PII) is exposed online and is at risk of identity theft and cybercrimes. The Identity Ecosystem developed by the Center for Identity in the University of Texas at Austin addresses this problem and provides a statistical framework for understanding the value, risk and mutual relationships of PII. The Identity Ecosystem currently uses a general Bayesian Network Model to simulate the relationships among PII, which may be quite inaccurate for specific groups of people. This thesis proposes a solution that specializes the Bayesian Network used for particular groups of people. Both one-dimension specialization and multi-dimension specialization are investigated. Research problems like how to choose specialization criterion, how to set specialization boundaries, and how to overcome the difficult of insufficient data, are carefully studied. Specialization functionality is demonstrated based on empirical data. Finally, experiments of specialization are conducted on data obtained from online stories. This work is important in the sense that it provides a guide-line of designing more accurate models of PII within the Identity Ecosystem.

List of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Cyberspace has changed our daily lives dramatically in the past several decades, including shopping, banking, accessing media, social networking, accessing company data, etc. Because of mobile technologies, this change is accelerated. It is reported that the number of newly activated Android devices is estimated to exceed 700,000 devices per day, which is almost twice the estimated growth rate of human population [1] [2]. The online availability of these services has resulted in greater opportunities for innovation and economic growth, but the infrastructure for supporting these services has not evolved at the same rate. This puts devices, individuals, businesses and governments at risk of identity theft and cybercrimes. Federal Trade Commission (FTC) in 2006 estimated an annual loss of over 15 billion dollars from identity theft [3]. In 2010 this figure had more than doubled, as 8.1 million U.S. adults were the victims of identity theft or fraud, with total costs of $37 billion [4]. Identity theft, according to the National Institute of Justice, has become the prime crime in the information age, with an estimated 9 million or more incidents each year [5].

Personally identifiable information (PII) needs to be better understood and valued for the purpose of security. The cyber world has merged into our everyday physical world, making a person's identity a complex intermingling of their online and offline attributes. Online attributes is composed of one's social media accounts, online shopping patterns, passwords, email accounts and so on. Offline attributes are those related to the physical world such as bank accounts, credit and debit cards, social security number, finger print, blood type, etc. It is evident that a more comprehensive online identity framework is needed based on sound understanding of PII.

The Identity Ecosystem developed at the Center for Identity at the University of Texas at Austin addresses this problem and provides a statistical framework for understanding the value, risk and mutual relationships of PII [6]. The Center's Identity Ecosystem uses a Bayesian Network Model to simulate the relationships among PII for individuals. It uses Bayesian Network inferences to provide a better understanding of the risk for PII exposure, and to give insights for protecting PII. For example, assume an identity theft steals a bunch of your information, such as social security number, email address, and birthday, the Identity Ecosystem is able to tell what will be the increased risk to other. The current Identity Ecosystem is limited to a single general model that hypothesizes all individuals have the same PII. Thus results of very generic and in some cases not applicable. For example, a 30 years old man may heavily rely on a social network in his daily life, but a 5-years-old child may not have a social network account at all. As a result, prediction of the Center's Identity Ecosystem for a particular person may not be precise at all, although the general model works well for all people as a whole. To solve this problem, this thesis proposes a solution that specializes the graphic model used for particular groups of people. That is, use some criteria to group people such that people in the same group are similar and people from distinct groups are different. For each group, a particular graphic model is learned and applied for prediction.

The rest of this thesis is arranged as follows: Chapter 2 briefly introduces the Identity Ecosystem, including its UI design, basic functions and queries. Chapter 3 explains the idea of specialization in details, including definition, ideal model, trade off, technologies solving trade off, etc. Chapter 4 demonstrates the success of the concepts of specialization based on empirical data. Chapter 5 briefly explained the ITAP project [7] [8], from which real world data are obtained. Then introduces methods of learning graphic model with those real world data and presents experiments results on queries.

Chapter 6 concludes with a summary of the thesis work. Chapter 7 discussed some possible future work.

Chapter 2: Background

To better understand our work, background knowledge including Bayesian Network model, Identity Ecosystem as a general graphic model (without specialization), and its UI, basic functions and queries are introduced in this chapter.

2.1 BAYESIAN NETWORK

Bayesian Network is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG) [11]. Bayesian Network has several advantages that make it suitable for the Identity Ecosystem [9]. First, Bayesian network allows one to learn about causal relationships, which are natural and universal in relationships of PII. Causal relationships not only help us understand the problem domain, but also allow predictions in the presence of interventions. Second, Bayesian Network is able to handle incomplete data sets. This has great importance for the Identity Ecosystem because the latter currently relies on data from online stories (will discuss in section 5.1), which are quite incomplete and sometimes not that accurate. Finally, Bayesian networks together with Bayesian statistical techniques facilitate the combination of domain knowledge and data. This is helpful especially when the Identity Ecosystem varies in data completeness and scale.

2.1.1 Bayesian Network Model

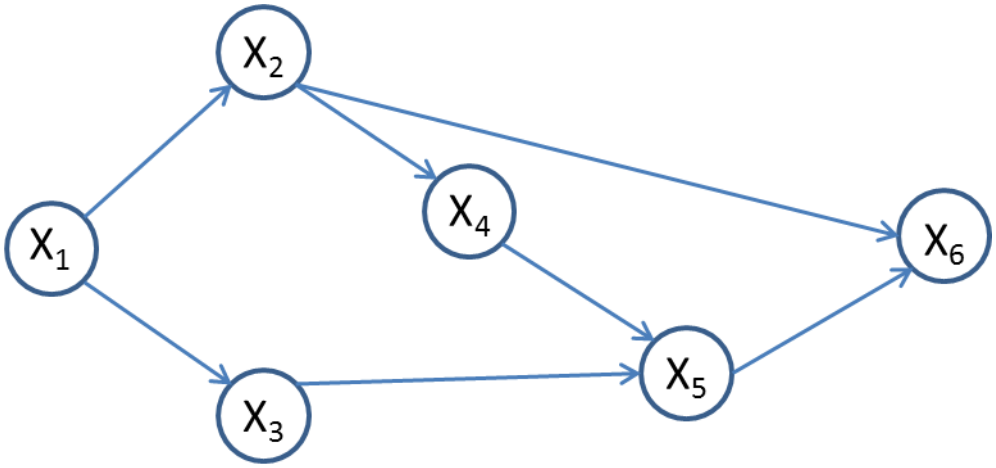

Figure 1: A simple example of Bayesian Network

An example of Bayesian Network is shown in Figure 1. It is a DAG (so as any other Bayesian Network) so that a jointed distribution $p(x_1, x_2, x_3, x_4, x_5, x_6)$ can be broken down into a product of conditional distributions following its topologic order, as shown in Equation 1.

$$p(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_4, x_3)p(x_6|x_2, x_5) \quad (1)$$

Equation 1 can be further simplified as shown in Equation 2, where pa_k denotes the set of parents of x_k.

$$p(x_1, x_2, x_3, x_4, x_5, x_6) = \prod_{k=1}^{6} p(x_k|pa_k) \quad (2)$$

Equation 2 can be generalized for any Bayesian Network as shown in Equation 3, where again pa_k denotes the set of parents of x_k, and $\mathbf{X} = \{x_1, \ldots\ldots\ldots, x_K\}$.

$$p(\mathbf{X}) = \prod_{i=1}^{K} p(x_k|pa_k) \quad (3)$$

The representation in Equation 3 implies that Bayesian Network significantly saves the amount of memory needed. Take the graph shown in Figure 1 for an example, a naïve way of storing the conditional probabilities of 6 binary variables needs storage

space of 2^6 values. However, by applying Equation 3, only 30 values are needed. This difference increases exponentially as the number of variables grows.

2.1.2 Probabilistic Inference

With all conditional probability distribution in a Bayesian Network, usually in the form of conditional probability table (CPT), one is able to calculate joint distribution $p(\mathbf{X})$ via Equation 3. With joint distribution $p(\mathbf{X})$, one can, in principle, inference any probability of interest. Typically, the desirable probability will be the probability distribution of unobserved variables given a set of variables (evidences). Unfortunately, the direct use of joint distribution $p(\mathbf{X})$ for inference becomes impractical when the number of variables is large [9]. There are two groups of practical inference methods for a general Bayesian Network, namely, exact inference methods and approximate inference methods. Exact inference methods include variable elimination (VE) [12], junction tree algorithm [13][14][15], recursive conditioning [11], AND/OR search [11], etc. All of those methods have time complexity exponentially in the network's treewidth, which is defined from the size of the largest clique in a chordal completion of the graph [16]. Exact inference thus is still quite expensive. Approximate inference tries to find a balance between speed and accuracy. It includes loopy belief propagation [17], generalized belief propagation [18], Variational methods [10], expectation propagation (EP) [19][20], sampling methods (also called Monte Carlo methods), and so on. Especially, sampling methods include various sampling ways such as rejection sampling, importance sampling [21], Gibbs sampling [22], etc. Identity Ecosystem applies Gibbs sampling because of its fast speed and its advantage of handling missing variables.

2.2 IDENTITY ECOSYSTEM OVERVIEW

As mentioned earlier in Chapter 1, Identity Ecosystem developed by the Center for Identity at the University of Texas at Austin provides a statistical framework for understanding the value, risk and mutual relationships of PII [6]. This section will briefly introduce the system.

2.2.1 Model Setup

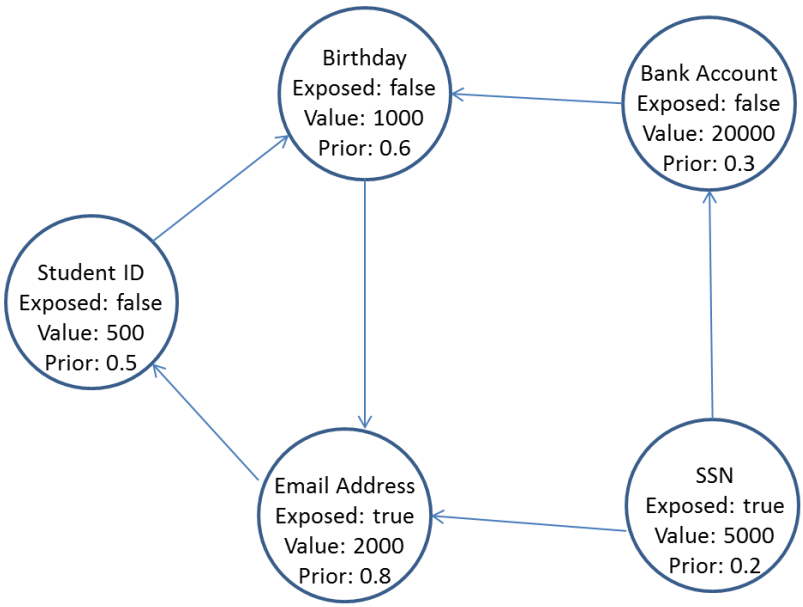

Figure 2: Mini example of Identity Ecosystem graphic model

The Bayesian Network is built such that each node x_k represents a type of PII, such as SSN, bank account, birthday, email address, etc. Each directed edge represents a causal relationship, which means the exposure of the start node may increase the probability of exposure of end node. A mini example of the Identity Ecosystem graphic model (IEGM) is shown in Figure 2. Each node has a Boolean flag, a value and a prior. The Boolean flag "exposed" denotes whether the node is exposed or not. The value

indicates the loss after the node is exposed. The value does not include any secondary loss, that is to say, it does not include the future loss because of the exposure of the current node. The prior denotes the probability of exposure of the node on its own. What is not shown in Figure 2 is a list of conditional probability estimates for each node x_k, representing how its parent nodes pa_k alone or in combination, impact the risk of the child node x_k (namely, the CPT of Bayesian Network) [6].

In fact, IEGM is much larger than the mini example. It has over 200 attributes (PII) from individuals, devices and companies. In this thesis, we mainly focus on attributes from individuals, whose number is 95 in total. IEGM is designed to be populated from two sources: expert knowledge and ITAP project. Details can be found in reference [6].

It has to be pointed out that the IEGM may have directed circles, for example, birthday -> email address -> student ID -> birthday in the mini example. Thus, IEGM is not a strict Bayesian Network. A solution of this is to use undirected graph instead, but this losses the causal relationship, which is important information. To overcome this problem, the circle is simply ignored when applying probabilistic inference. This strategy will make the inference less accurate. Fortunately, the actual model turns out to be sparse, which means the case of directed circle is rare.

2.2.2 UI and Functions

The main interface, the Identity Ecosystem Viewer, is a visualization tool of the Identity Ecosystem, and shown in Figure 3. The 3D graphic model is visualized in the screen and can be moved and rotated. The right most part is the control menu. “Filters” control the types of nodes and edges to be displayed. “Controls” contain two buttons:

"Display Attributes", which shows properties of a chosen attribute, and "Refresh Display", which refreshes the 3D graphic model. "Color/Size Options" allows users to choose a node property to determine node size and color in the 3D graphic Model. Details about node color and size are shown in the bottom right. The top left part includes two combo bars "Ask A Question" and "Specialization Charts" and a button "Specialization Options". "Ask A Question" provides three types of queries (will be explained in details in section 2.2.3). "Specialization Charts" provides statistic charts for specialization and "Specialization Options" allows the user to decide in which way the graphic model is specialized. Concepts of specialization will be discussed in Chapter 3.

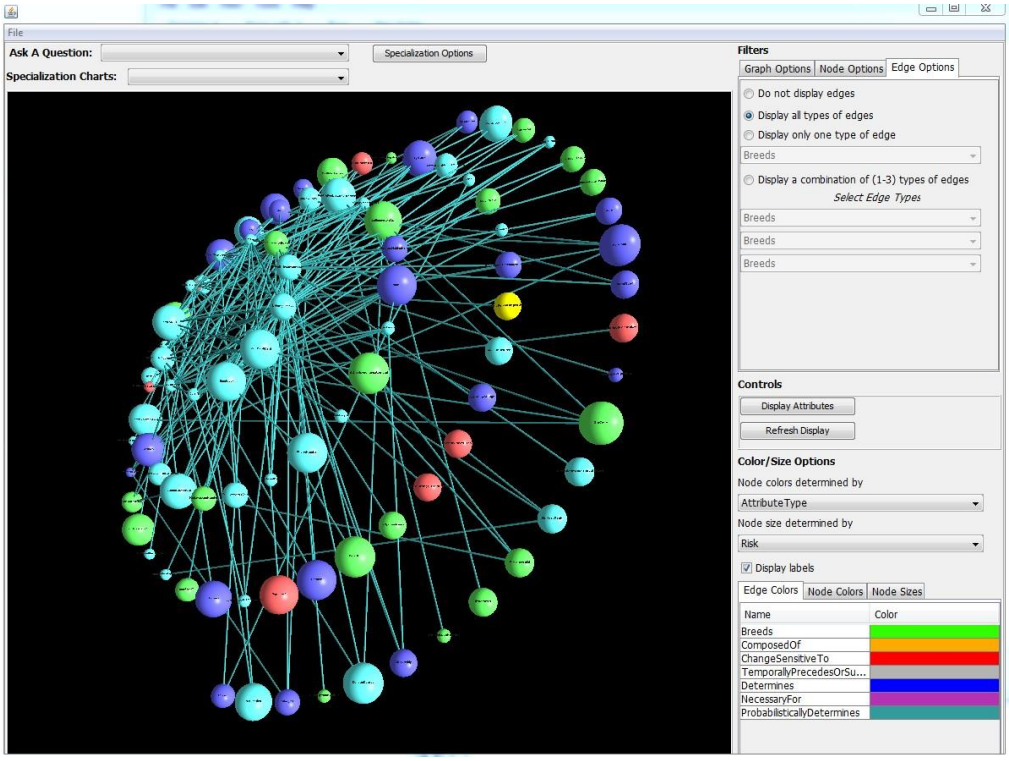

Figure 3: Main interface of Identity Ecosystem

2.2.3 Queries

Three types of queries (inferences) are supported in the Identity Ecosystem, namely, infer probability of breach based on evidence (PII Exposure Query), detect most probable origin of a breach (Breach Origin Query), and find breach hotspots (Hotspots Query). It is a good idea to illustrate these three types of queries by user stories.

For PII Exposure Query, imagine one day you click on a harmful website by mistake, which asks for your personal information for the purpose of registration. You type in your SSN before you realize it is dangerous. You'd like to find out, after the exposure of your SSN, what risk has been imposed on your other identity attributes as a result of the SSN exposure. Run the PII Exposure Query, in the interface shown in Figure 4, choose, SSN as evidence, and run the query.

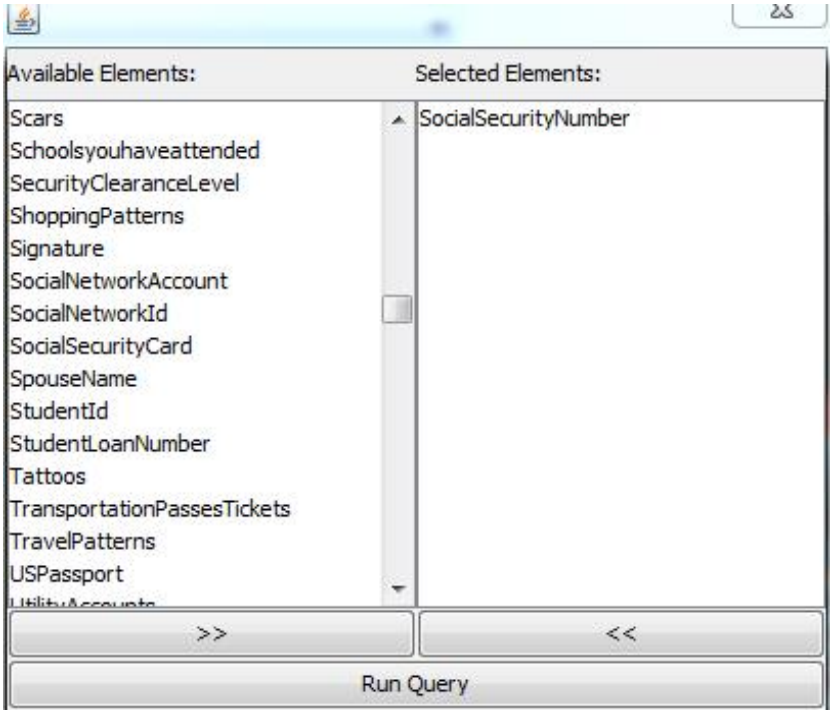

Figure 4: Interface for choosing attributes as evidences

After running the query, you will get the results represented in two different ways. Figure 5 shows the results of the query in a bar graph, in which every affected node has a value indicating the increase of its potential loss given the breach of SSN. It indicates that your Bank Account and Credit Card Number are at highest risk. You probably want to take action to prevent such loss. Figure 6 shows the results within the 3D graphic model. Red color with large size represents high risk, yellow color with medium size means medium risk, white color with small size denotes low risk, and orange color with small size represents unaffected. Bank Account and Credit Card Number are marked as red.

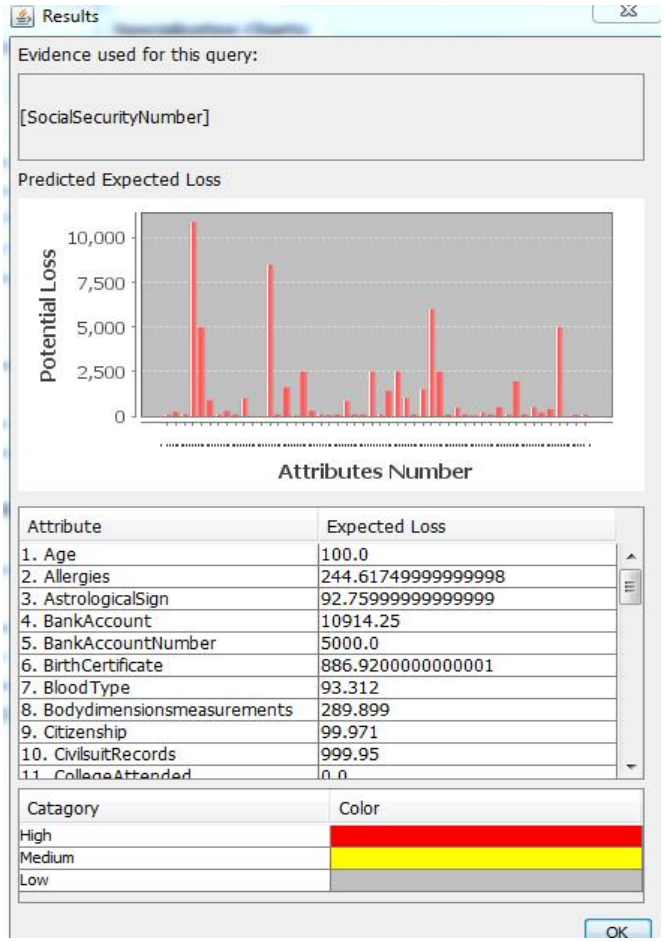

Figure 5: Results of the risks due to PII Exposure Query represented in bar graph

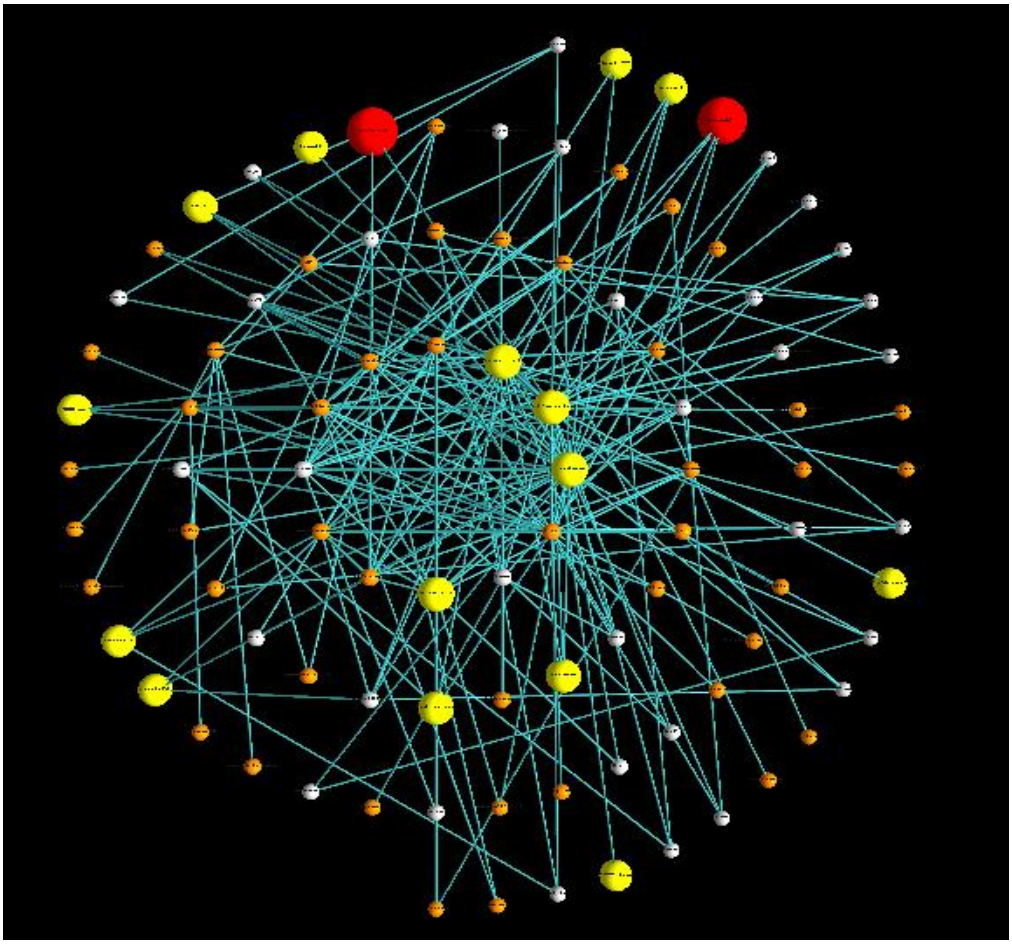

Figure 6: Results of the risks due to PII Exposure Query represented in 3D graph

For the Breach Origin Query, assume suddenly, there are many more junk emails sent to your email every single day. This implies that your email account is exposed. You want to figure out what is the most probable origin of this breach so that you can prevent such a thing happening in the future. Run the Breach Origin Query, select Email Account as the evidence and run the query. You get a 3D graph and a bar graph showing the most probable source is Social Network Account. You probably need to change your password to prevent further loss. Here to avoid repeated work, only the bar graph is shown in Figure 7.

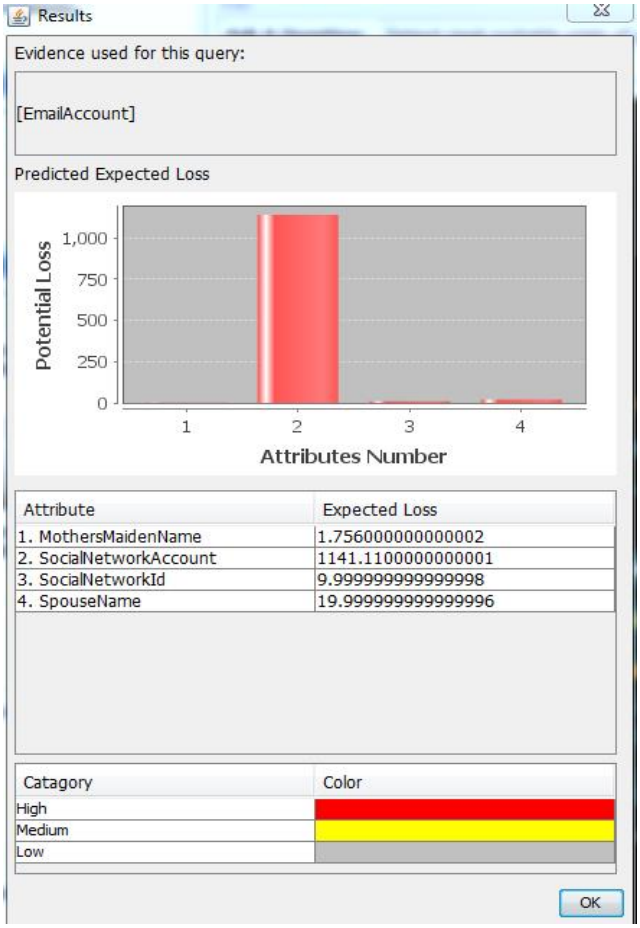

Figure 7: Results of the Breach Origin Query represented in bar graph

For the Hotspots Query, we continue to use the story describing a breached email account. You want to find the breach hotspots so you can prevent further breaches (the nodes whose exposure will cost the most in terms of total loss: intrinsic loss plus secondary loss downstream). A 3D graph and a bar graph are shown immediately after you run the query. These indicate that your Bank Account, Credit Card Number, Credit Debit Card and DNA are hotspots. Figure 8 shows the results of query in a bar graph.

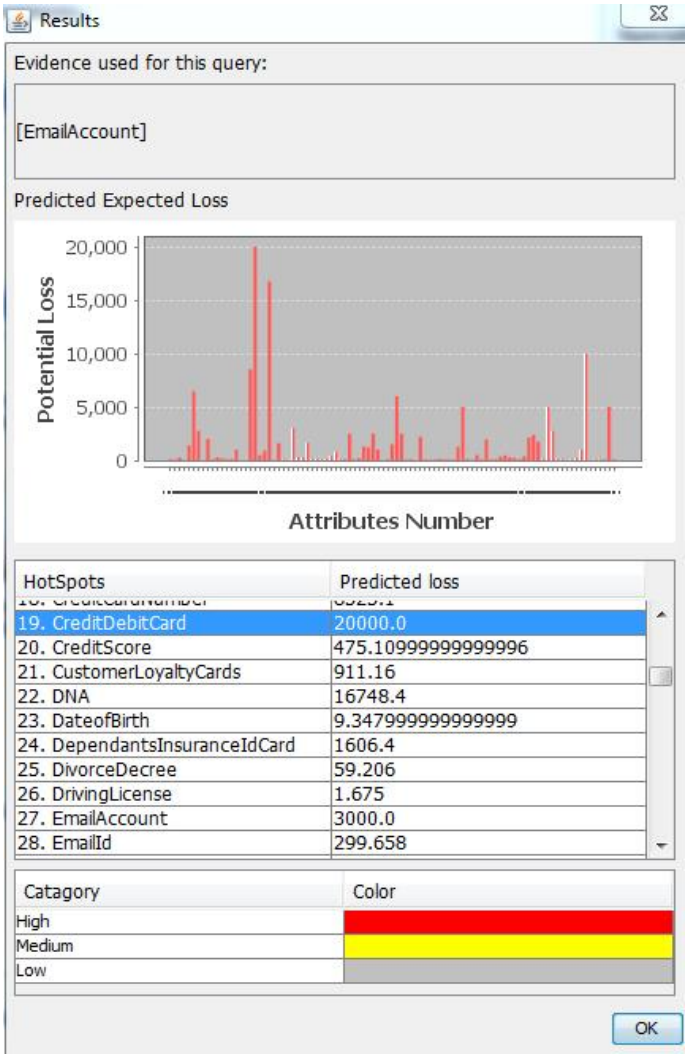

Figure 8: Results of the Hotspots Query represented in bar graph

It is important to note that the graphic model used here for the three types of queries using empirical data obtained from subject matter experts. Currently the Ecosystem values of PII attribute risks and values are obtained from the Center Identity Threat Assessment and Prediction (ITAP). The ITAP project is currently collecting data about identity theft and fraud crimes to ascertain criminal behaviors and consequently risks of exposure to PII and the values/benefits criminals reap from PII.

Algorithms used for the queries are Bayesian Network inferences, which use Gibbs sampling as the sampling method [10][22]. Especially, the Breach Origin Query used junction tree algorithm to find most probable source of a breach [10][13]. Details about algorithms applied can be found in the reference 6 [6].

Chapter 3: Theories and Technologies of Specialization

As mentioned earlier in the Introduction, prediction based on a general graphic model may not be accurate enough for all people. In fact, a prediction may be far from the truth in some particular cases. For example, given one's SSN is exposed, the general graphic model may predict that his or her bank account is at high risk with potential loss of about $10000, while actually one has little income and has no money in his or her bank account. The purpose of this thesis is to improve the accuracy of prediction by specialization. This chapter discusses specialization of Identity Ecosystem. Topics include definition of specialization, specialization methods, trade off and solutions, evaluating a specialization, and multi-dimension specialization.

3.1 Definition of Specialization

The term "specialization" or its variances was used in many disciplines [23]. In academia, "academic specialization" means a course or major or may refer to a field a specialist practices in. In biology, terms like "cellular differentiation" is the process by which a less specialized cell becomes a more specialized cell type. In computer science, "template specialization" indicates a style of computer programming which allows alterative implementations to be provided based on certain characteristics of the parameterized type that is being instantiated. While in Economics and industry, "Specialization (functional)" represents the separation of tasks within a system. Although "specialization" has different meanings in distinct disciplines, there are two points in common:

1. Each specialized item aims to a smaller scope of application.
2. Compared with general items, specialized ones are specific or precise scope.

Based on the two points, the definition of specialization in the Identity Ecosystem is proposed to be: developing specialized models for different groups of people such that each specialized model works better than the general model for a specific group of people.

3.2 PROPOSING SPECIALIZATION LIST

Before proposing suitable ways of specializing the Identity Ecosystem graphic model, one may ask two questions: how the technology "specialization" really works? And how well a certain way of specialization is? This section tries to answer the two questions and to propose suitable ways of specialization.

3.2.1 How Specialization Works

Figure 9 attempts to illustrate how specialization works. For an arbitrary attribute, its real values for different people (different samples) have a distribution shown in Figure 9 (a). A general graphic model simply uses the average of those values as the attribute value, which may have large errors for specific people. Instead, specialization tries to use extra information to group people (we call this extra information "specialization criterion"). The example shown in Figure 9 (b) indicates that, with specialization criterion "age", the attribute values can be well grouped into two groups. One group is under 25 years old, represented by green dots, the other is over 25 years old, represented by red stars. Each of the group has a smaller range of attribute values, and, therefore, specialized graphic models based on average attribute values of these specialized groups are more accurate than the general model for their corresponding specialized groups. Specialization criterion can be many things like gender, income, location, etc. as well as age (details will be shown in section 3.2.3). Note that data used in the example shown in

Figure 9 is not real data. In most cases, the effect of specialization may not be that obvious, but the example does illustrate the basic idea how specialization works. In addition, a typical specialization has more than two specialized groups.

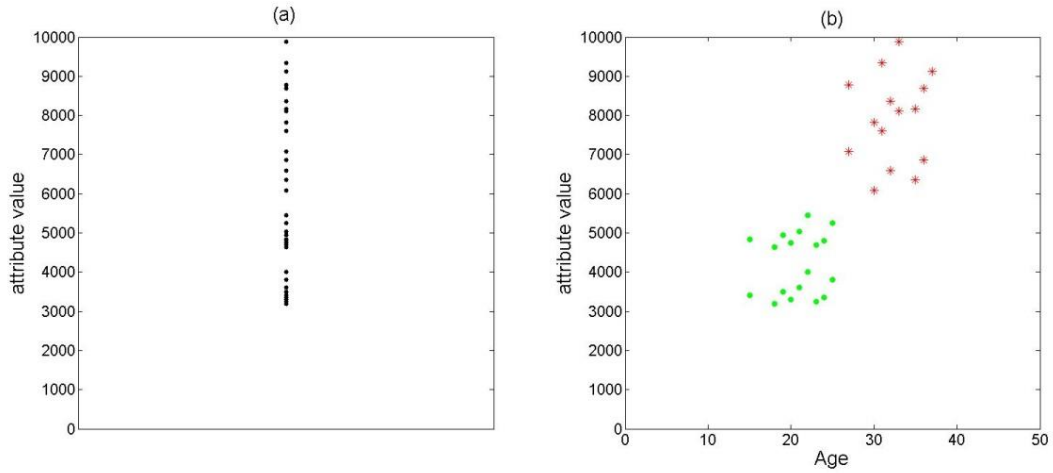

Figure 9： (a) Distribution of attribute values from different people with no extra dimension (b) distribution of attribute values from different people with an extra dimension: “Age”

3.2.2 Evaluation of Specialization

Specialization is similar to a classification problem. The final goal is to group people such that, with a suitable number of specialized groups, variation of attribute values in each specialized group is minimized and differences among distinct specialized groups are maximized. Evaluation of specialization criterion is related to the extent such a criterion achieves the goal. Remember that evaluation of a particular specialization criterion is meaningful only when the amount of data available is limited and the number of specialized groups is bounded. With unlimited amount of data sample for each person,

one can always achieve the best specialization by simply modeling each single person as a specialized group.

For the purpose of evaluating specialization, we can borrow the idea of a classic classification algorithm: Fisher's Linear Discriminant Analysis (FLDA) [10]. For a two-class classification problem in a K-dimension space, FLDA uses a K-1 dimension hyperplane (shown in Equation 4) to define the boundary between class C_1 and C_2.

$$y = \mathbf{A}^T\mathbf{x} + A_0 \tag{4}$$

Assume with the hyperplane defined by Equation 4, we obtained N_1 vectors in C_1 and N_{21} vectors in C_2. The mean vectors of the two classes are given by Equation 5 and 6:

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{k\in C_1}\mathbf{x}_k \tag{5}$$

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{k\in C_1}\mathbf{x}_k \tag{6}$$

The goal is to find a vector **A** such that the metric $M(\mathbf{A})$ (defined in Equation 7) is maximized.

$$M(\mathbf{A}) = \frac{(m_2-m_1)^2}{s_1^2+s_2^2} \tag{7}$$

where $m_2 - m_1 = \mathbf{A}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)$ is the separation of two class means projected on the hyperplane defined by equation 4, and s_1^2 and s_2^2 are within-class variance of the projected data from C_1 and C_2, respectively.

In the case of specialization, the space is two dimensional: attributes value and specialization criterion used (for the example in Figure 9, it is "age"). Since in a two-class two-dimension problem, the vector **A** in Equation 7 becomes a one-dimension variable A, which represents the slope of the boundary line. One can similarly define the metric for a two-group specialization problem to be Equation 8:

$$M(k) = \frac{(m_2-m_1)^2}{s_1^2+s_2^2} \tag{8}$$

The difference is that the boundary for the two groups is a line parallel to the axis of "attribute value" (for example, age = 25 in Figure 9) and k is a variable instead of a vector that determines the position of the boundary. In the same way, $m_2 - m_1$ is the separation of two class means projected on the boundary, and s_1^2 and s_2^2 are within-class variance of the projected data from C_1 and C_2, respectively.

Extending this idea to a multi-group specialization problem is tricky. When FLDA is extended to multi-class problem, at most K-1 linear 'features' can be found with K-dimension [10] [24], which means extension of Equation 7 for FLDA to multi-class problem can only succeed when class number is no more than the number of dimensions. Unfortunately, the specialization problem discussed here has only two dimensions. Therefore, Equation 8, for the same reason, cannot be directly extended for a multi-group specialization problem. Here an alternative evaluation equation for a K-group specialization problem is proposed as Equation 9:

$$\mathrm{M}(\mathbf{k}) = \frac{1}{\sum_{i=1}^{K-1} \frac{s_{i+1}^2 + s_i^2}{(m_{i+1} - m_i)^2}} \tag{9}$$

Where **k** is a (K-1)-dimension vector that determine the boundary of K groups. m_i and s_i^2 are mean and variance of projected value in class C_i, respectively. m_i is indexed such that $m_i \geq m_j$ if $\mathrm{i} > \mathrm{j}$. The strict proof defining why Equation 9 makes sense has not been completed. However, it makes sense in the way that when maximizing $\mathrm{M}(\mathbf{k})$, all differences between two adjacent means $m_{i+1} - m_i$ should be maximized while the sum of their variances $s_{i+1}^2 + s_i^2$ should be minimized.

Consequently, maximizing a specialization metric is actually the process of optimizing boundaries. The score for a specialization criterion can be defined as the maximum metric it can achieve with a given group number. Given a fixed group number,

the way that has a higher metric is the better one. Note that two different specialization criterion with different group number cannot be compared directly by their metrics.

An implicit assumption for the definition of metric in Equation 7-9 is that the sample points in each class follow Gaussian distribution. One of the challenges in this work is lacking in data. Therefore, the exact distribution data points follow is unknown. Gaussian distribution is, in theory, the most probable distribution, because of random variables [10]. It is still possible that the data points follow other kinds of distribution. It is also possible that with different methods of specialization, the distribution patterns are distinct. For other distributions, one needs to replace variances in Equation 7-9 by other measurements. For example, if data points follow linear distribution, variance within C_j should be replaced by $\sum_i |x_{ij} - \mu_j|$, where x_{ij} is a projected data point in class C_j and μ_j is the projected mean in C_j.

3.2.3 Specialization List

Table 1: Proposed specialization list

Criterion	Group Name					
Age	Baby	Child	Teenager	Adult	Senior	Deceased
Gender	Male	Female				
Education Level	Uneducated	Primary School	Middle School	High School	College	Graduate School
Profession	Engineering	Business	Science	Medical	Management	Others
Income (annual)	<10k	10-20k	20-50k	50-100k	100-500k	>500k
Location	East	West	South	North	Middle	Aboard
Citizenship	North America	South America	Europe	Asian	Africa	Australian

Table 1 outlines the list of specialization criteria. These specialization criteria are abstractions suggested by the current attributes held in the Identity Ecosystem. "Age" makes sense because people within different age groups may have huge difference in their life styles, PII they have, etc. For example, a 20-years-old college student may heavily rely on a social network like Facebook, but a baby may not have social network accounts at all. An obvious effect from "Gender" may be shopping patterns. Women may like makeup and new clothes, but men may prefer video games. "Income" directly determines how much one may lose after a certain PII is exposed. In the same way, one can justify the other proposed specialization criteria.

Boundaries for the proposed specialization criteria are estimated. A big challenge for this thesis is lacking of real data, so optimized boundaries are not able to be obtained at this current stage. The experimental data relies on the ITAP program [7][8], which is currently manually collecting data from reports of identity theft and fraud. However, enough data will be available in the near future, allowing one to redefine boundaries based on data using methods discussed in section 3.2.2.

Finally, the proposed specialization list and boundaries are for all attributes for individuals used in Identity Ecosystem, not just for one particular attribute. In other words, all attributes share the same specialization boundaries in this thesis. This means we need to maximize M, the sum (or weighted sum) of metrics of all specialized attributes, using Equation 10 while trying to find optimized boundaries.

$$\mathrm{M} = \sum_{i=1}^{N} \mathrm{M}_i(\mathbf{k}_i) \tag{10}$$

Where N is the number of attributes specialized and $\mathrm{M}_i(\mathbf{k}_i)$ is metric of the i^{th} specialized attribute.

3.3 MULTI-DIMENSION SPECIALIZATION

Continue the example shown in Figure 9, in which we add another dimension "income". The new distribution of sample points is shown in Figure 10. Now one can group the attribute values by a combination of "age" and "income". Group1 has age < 25 and income < 50k, denoted by green squares, group 2 has age < 25 and income > 50 k, marked by blue circles, group 3 has age > 25 and income < 50k, represented by red dots, and group 4 has age >25 and income > 50k, noted by blue stars. Each multi-dimension specialized group has a smaller range of attribute values than that in the general model as well as in corresponding one-dimension specialized models.

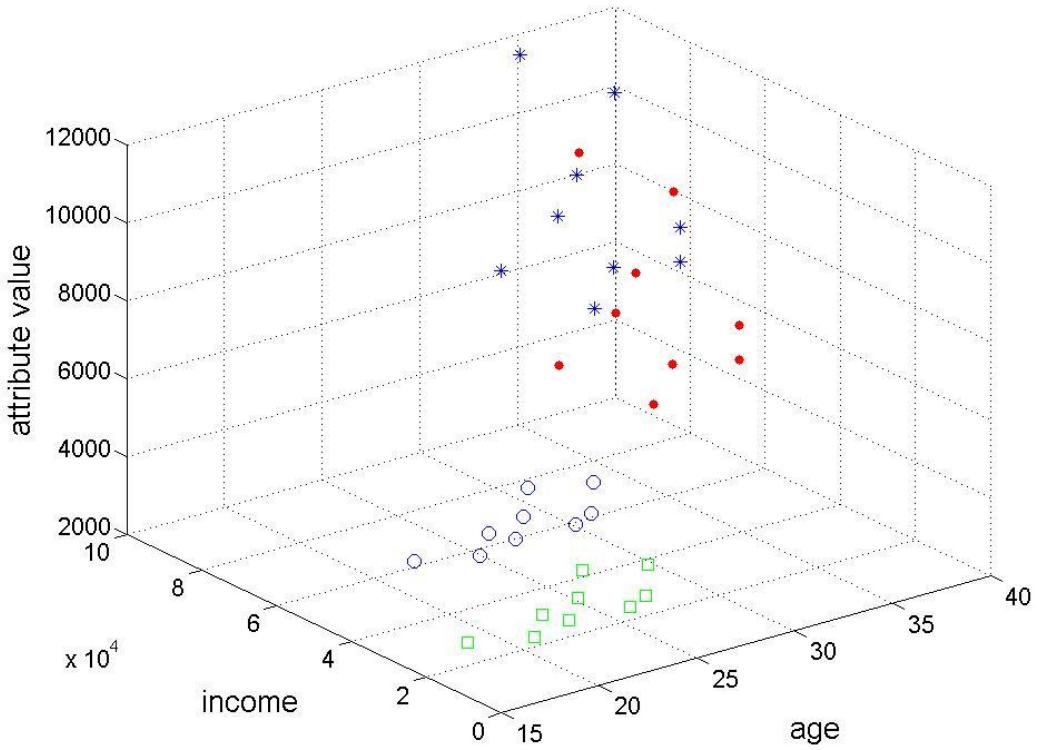

Figure 10: Distribution of attribute values from different people with two extra dimensions: "Age" and "Income"

Metrics for multi-dimension specialization can be obtained by expanding Equation 9 to a multi-dimension space, from which optimized boundaries can be determined. Details are not discussed here.

Here linear boundaries, or more precisely, grids, are used, which is the weakest way of classification. For multi-dimension specialization, it is possible to use different types of boundaries in the multi-dimension space except the axis "attribute value". For example, in the example shown in Figure 10, it is possible to apply other types of boundary in the two-dimension space composed of "age" and "income". Those boundaries can be general lines (with a slope other than lines parallel to the axis "attribute value") provided by classifiers like FLDA [10], or some nonlinear boundaries determined by unsupervised classifiers such as K-mean [25]. This thesis focuses on "grid" boundaries, since not enough data is available for analyzing more complicated boundaries.

3.4 IDEAL CASE & TRADE OFF

Perhaps the ideal case for all machine learning problems occurs when there are unlimited amounts of data. Under such condition, the best specialization would be building a specialized model for each person. Unfortunately, it seems impossible for a single person to have enough records to populate such a large Bayesian Network. A more practical case is to use multi-dimension specialization with all specialization criteria proposed in Table 1, in which each specialized model, on average, has an amount of data that is $2 \times 6^6 = 93312$ (gender has two groups and all other six PII have six groups each) times smaller than the total amount of data available. Let's say each specialized model needs about 1000 data points to achieve a reasonable performance, the total

number of data points needed is at least 93.3 million. This is still a mission impossible for our ITAP program [7] [8] to date.

Define the degree of specialization to be as Equation 11:

$$\mathrm{D} = \prod_{i=1}^{N} S_i \tag{11}$$

where N is the number of specialization criteria used for specialization and S_i is the number of groups for the i^{th} criterion used. For example, in Figure 10 the degree of specialization is $\mathrm{D} = 2 \times 2 = 4$. Obviously, the degree of specialization is equivalent to the number of specialized groups under a multi-dimension specialization.

The trade-off is: with a larger degree D, scope of each specialized group is smaller, thus a specialized model may be more accurate for its corresponding specialized group given enough data. On the other hand, the average amount of data available for each specialized model is the total amount of data divided by the specialization degree D. Thus less data can be used to train a specialized model with larger degree D, which may make the specialized model less accurate. Given certain amount of training data, an optimal value of D must exist.

3.5 SOLUTIONS

Section 3.4 shows that there is, in theory, an optimized degree of specialization D, given certain amount of training data. At the time this thesis is written, the amount of data available is quite limited, 791 online stories from ITAP project. Thus, the best strategy would be to eliminate as much unnecessary specialization as possible. In this section, several technologies minimizing unnecessary specialization are proposed.

3.5.1 Eliminating Correlation

One may already notice that some of the PII used for specialization in Table 1 may be highly correlated. For example, one with higher education is likely to have high annual income. Also, babies and children are likely to be uneducated or in primary schools, and have little income. With limited amount of data, it is not wise to use correlated specialization criteria such as "education" and "income". Steps of choosing ways of specialization criterion are proposed as:

1. Figure out the total number of data points available, let's say N, and N_s, an average number of data points needed for a specialized model to work appropriately.
2. Calculate the maximum degree of specialization acceptable: $D_{max} = \frac{N}{N_s}$
3. Choose a combination of specialization criteria and group number of each criterion, with total degree no more than D_{max}. Make sure specialization criteria used provides effective specialization and correlation among them is minimized.

3.5.2 Partial Specialization

The idea of partial specialization is that: given so many variables and parameters in a Bayesian Network, it is not necessary to specialize all of them. That is to say, different specialized models can share some variables or parameters, and, therefore, share training data for those shared variables and parameters. This section discusses two ways of partial specialization.

3.5.2.1 Partial Specialization on Attributes

After carefully examining all attributes in the Identity Ecosystem with all proposed specialization criteria in Table 1, the research found that a particular specialization criterion may not make sense for all attributes. For example, the criterion "age" has nothing to do with the attribute "name". No matter what age one is, his or her name is in the same risk. The criterion "gender", for another example, does not affect the attribute "divorce decree". After all, man or woman cannot divorce by himself or herself.

Table 2 shows the statistic results of the number of attributes that will be unnecessarily specialized by each specialization criterion. From it, we know that, "Age" is the most efficient criterion for specialization, with a percentage of attributes unnecessary for specialization as low as 11.6%. This makes sense because "Age" groups influence one's life style significantly. "Profession" turns out to be the least efficient criterion, with a percentage of 89.5%. On average, about half of the attributes do not need to be specialized by applying partial specialization on attributes. This implies that much more data points can be used to learn parameters related to those unspecialized attributes.

Table 2: Statistic results of unnecessary specialization on attributes

Criterion	Age	Gender	Education	Profession	Income	Location	Citizenship
No. of unnecessary attributes	11	56	44	85	34	50	66
Percentage (%)	11.6	58.9	46.3	89.5	35.8	52.6	69.5

3.5.2.2 Partial Specialization on Parameters

There are three types of parameters in Identity Ecosystem graphic model, namely, conditional probability table (CPT), attribute values, and attribute priors (intrinsic probability that an attribute is exposed). Attribute priors can be either set based on prior knowledge or learned from training data in some way. If they are set based on prior knowledge, then specialization is straightforward: simply set different values to priors in different specialized models according to prior knowledge. CPT and attribute values need to be learned from training data. For a Bayesian Network with 95 nodes like in the Identity Ecosystem, there will be 95 attribute values. However, the number of conditional probability depends on the structure of the Bayesian Network, and is probably much larger than 95. Assume that each node has an average number of causal links of 5, so the number of total conditional probability becomes $95 \times 2^5 = 3295$. For a dense graph, this number grows exponentially. This fact implies that given the same requirement of accuracy, learning CPT needs much more data than learning attribute values. Therefore, one can merely specialize attribute values, and allow different specialized models to share a common CPT, if the amount of data available is limited. This strategy makes sense for another reason. In common sense, the difference of conditional probability for different people may be much smaller than that of attribute values, since identity thieves apply similar approaches of getting information, but how much such information is worth highly depends on who this piece of information belongs to.

The strategy discussion above works for both one-dimension specialization and multi-dimension specialization. If more training data were available, however, an alternative technology may be applied to multi-dimension specialization: If one has enough data to populate a CPT for one-dimension specialized models, but not enough for multi-dimension specialization, one can simply specialize the CPT for each one-

dimension specialization and estimate the CPT of a multi-dimension specialization based on those one-dimension specialized CPTs. Here the research proposes a way of estimating the CPT of a multi-dimension specialization: simply use the average of conditional probabilities in each one-dimension specialized CPT involved. For example, a conditional probability in an "age-income" two-dimension specialization can be the average of corresponding conditional probabilities in "age alone" specialization and "income alone" specialization.

Similarly, if training data is enough to train attribute values for one-dimension specialization, but not enough for multi-dimension specialization, one can simple use the average of attributes values from all corresponding one-dimension specialization models as the attribute value of the multi-dimension specialization model.

Chapter 4: Demonstration of Specialization based on Empirical Data

Due to the limited amount of data available from ITAP at the current stage, demonstration of specialization based on real data is challenging. Attempts are shown in Chapter 5. Alternatively, this chapter focuses on demonstrating the functionality of specialization in Identity Ecosystem based on Empirical data. Empirical data, although not as trustworthy as real data, populate a complete graphic model (every parameter in the model is populated). Therefore, it is currently the best way to show the impact of specialization on risk prediction.

4.1 SPECIALIZING EMPIRICAL DATA

In section 2.2.3, empirical data is already used for demonstration of the three types of queries. The data include CPT, attribute values, and attribute priors. For the purpose of demonstrating impact of specialization, partial specialization on the Identity Ecosystem by only specialized attribute priors is enough. For each specialization criterion listed in Table 1, a file includes all attribute priors for each specialized group in the criterion is generated based on prior knowledge. These data can be directly applied for demonstration of one-dimension specialization. For multi-dimension specialization, an average prior for all one-dimension specialized priors is applied for each attribute (the same technology as discussed in section 3.5.2.2 for estimating multi-dimensional specialized CPT).

4.2 DEMONSTRATION

Remember that in section 2.2.2 the main interface of Identity Ecosystem (shown in Figure 3) is introduced. At the top left side of the interface, there is a combo bar

"Specialization Charts" for specialization related charts and a button "Specialization Options" to set specialization criteria for a specialized model. This section demonstrates these two functionalities.

4.2.1 Specialization Charts

The first function is to visualize specialized attribute priors with their grades: high, medium, low, and no risk, instead of actual values. Take "age" as the specialization criterion for example, first click on the combo bar "Specialization Charts" and choose "risk per age". Then the user is asked to choose attributes to be visualized in the same way as shown in Figure 4. Finally the user runs the query and the results are shown. Figure 11 shows the results with attributes: bank account, blood type, and email account, where red means high risk, yellow represents medium risk, green denotes low risk and blue means ignorable risk. The results make sense in the way that baby has smaller risk of the three attributes than almost all other age because PII may have not been initialized for them.

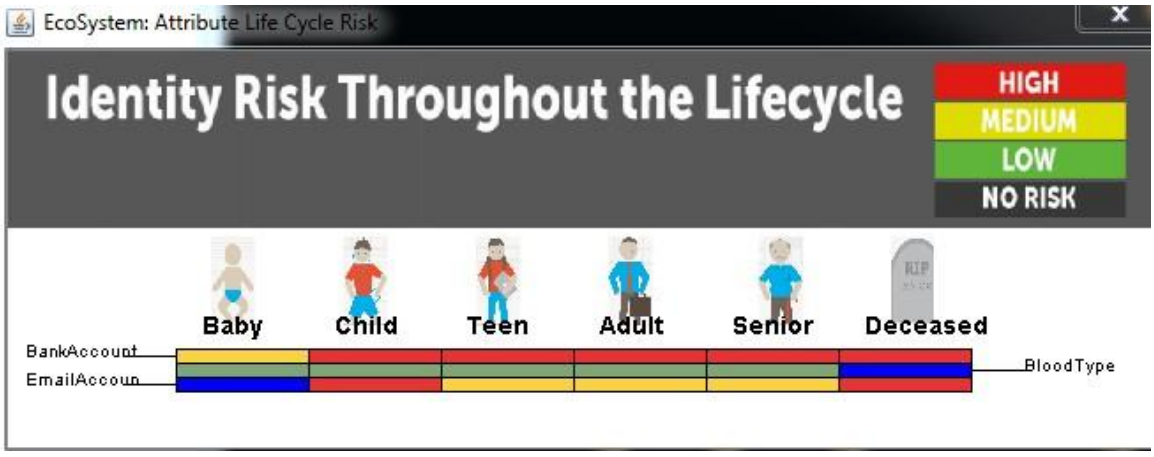

Figure 11: Risk per age visualization

Another function developed for the Identity Ecosystem is risk chart, which plots based on exact attribute priors. This time, the use selects “income” as the specialization criterion. The user clicks on the combo bar “Specialization Charts” and chooses “income risk chart”, and follows the same steps for choosing attributes. Results are shown in Figure 12 with four attributes: bank account number, college attended, email Id, and vehicle registration license plate. Take “bank account number” for example, it makes sense that one with more income, is likely to have more bank accounts and use them more frequently, so that his or her bank account number is more likely to be exposed.

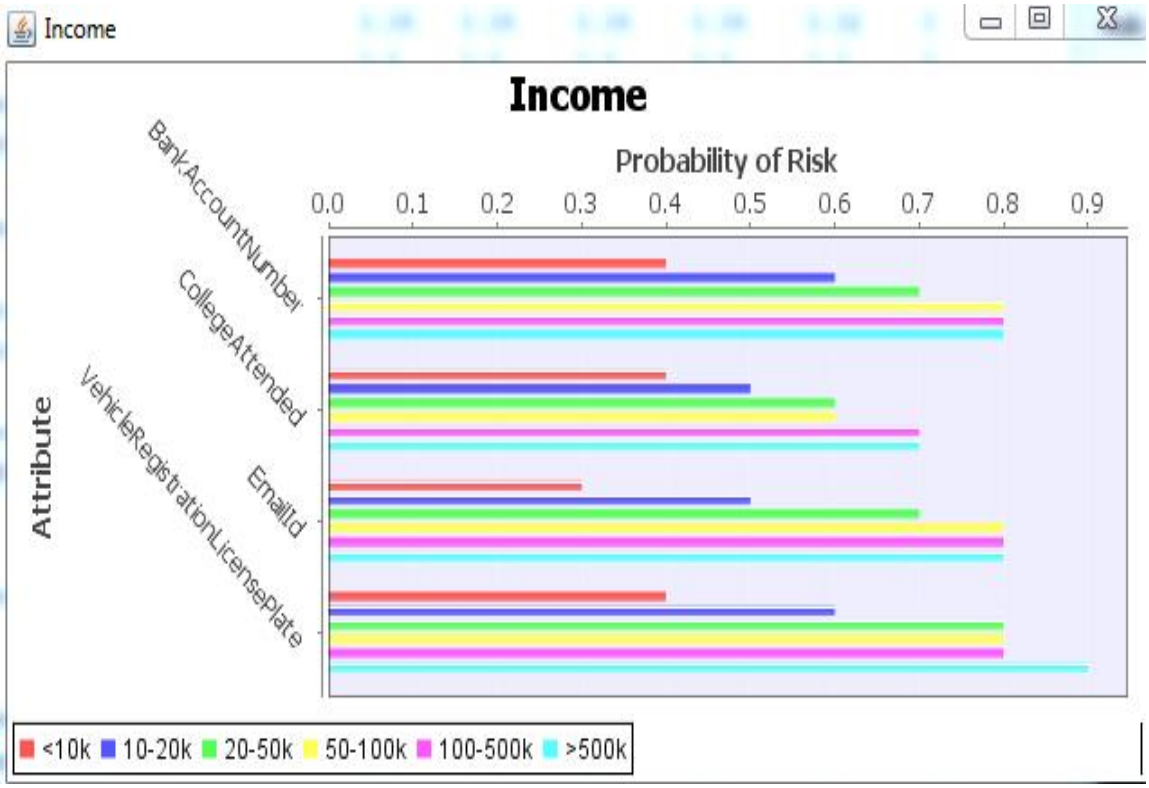

Figure 12: Risk Vs income chart

4.2.2 Queries in Specialized Models

To set specialization for queries, click on button “Specialization Options” shown in Figure 3, a control interface popups as shown in Figure 13. The “Model Setting” has two options: “Single Choice”, which is for one-dimension specialization, and “Multiple Choices”, which is for multi-dimension specialization. In “Single Choice” model, only one specialization criterion can be chosen, while in “Multiple Choices”, at least one criterion can be chosen. “Reset” button clears all previous choices. After setting, click “OK” and now the three queries will be conduct based on the specialized model chosen.

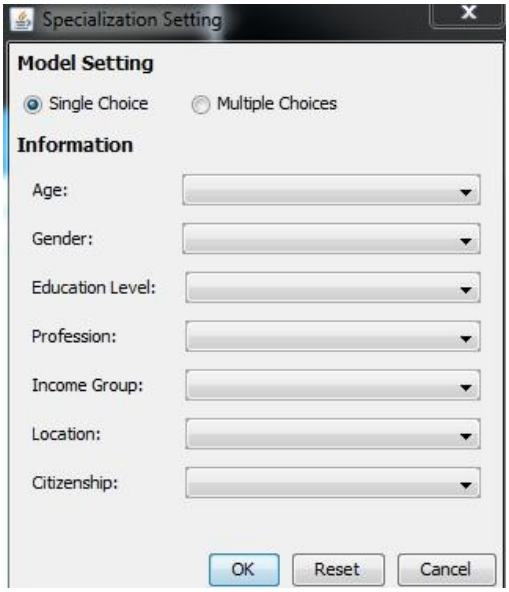

Figure 13: Interface for specialization setting

To avoid repeat work, the following shows the demonstration of one-dimension specialization, only demonstrate PII Exposure Query discussed in section 2.2.3. For the demonstration, the user selects “age” as specialization criterion and select “child” as specialized group. Again, the user uses “Social Security Number” as evidence and run PII Exposure Query, the results as a bar graph is shown in Figure 14. Compared with the

results from the general model shown in Figure 5, the results here are quite different. For example, bank account has a potential loss of around 7000, much smaller than that in Figure 5. This makes sense because child may have less money than average case. However, 7000 may still sound too high for a child, this is because we merely specialized attribute priors, but use the same attribute value for bank account. A more complete, pervasive specialization across the attributes will make the prediction more accurate.

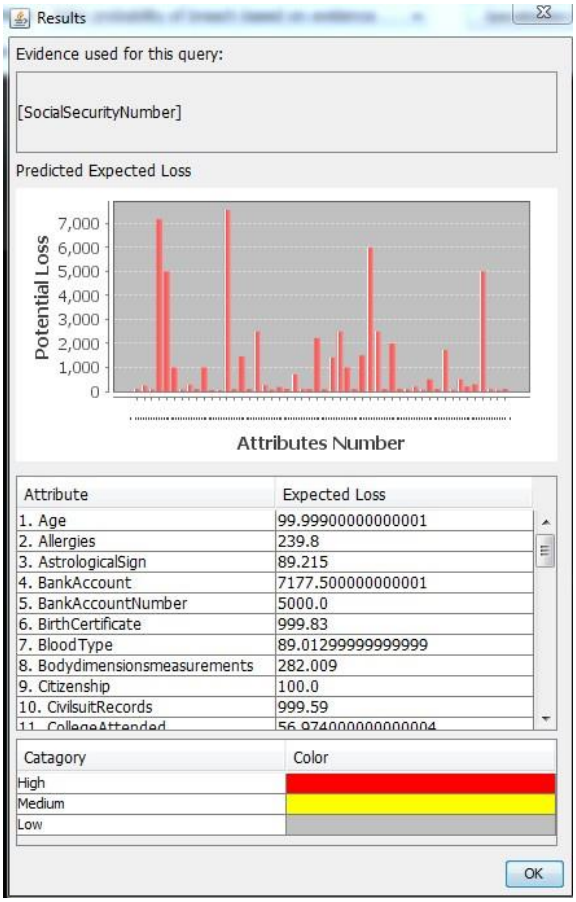

Figure 14: Results of PII Exposure Query in bar graph, with one-dimension specialization "age-child"

The results as a 3D graph are shown in Figure 15. There are 5 attributes marked in red: bank account, bank account number, credit card number, medical history and social security number, more than that shown in Figure 6. Note that "high" risk is a relative concept, defined as more than 75% of the highest one.

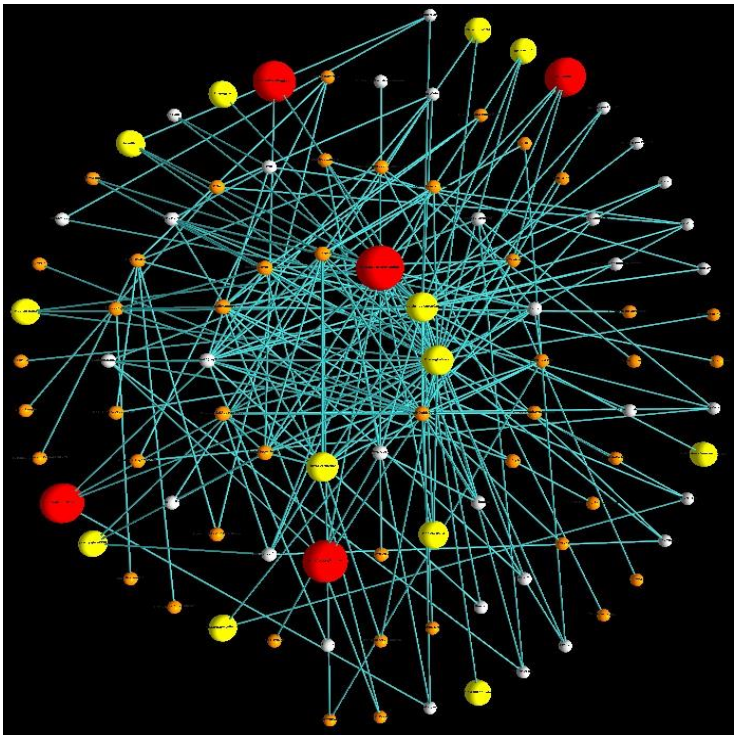

Figure 15: Results of PII Exposure Query Represented in 3D Graph, with One-Dimension Specialization "Age-Child"

For multi-dimension specialization, the user issue the hotspots query discussed in section 2.2.3. Specialization criteria and corresponding groups are: Age & Adult, Gender & Male, Income & 10-20k, and Location & Aboard. The results as a bar graph are shown in Figure 16. Compared with results from general model in Figure 8, the overall potential

loss reduced, which make sense because the example used has low income. Fraction of loss for each attribute also changed because of specialization.

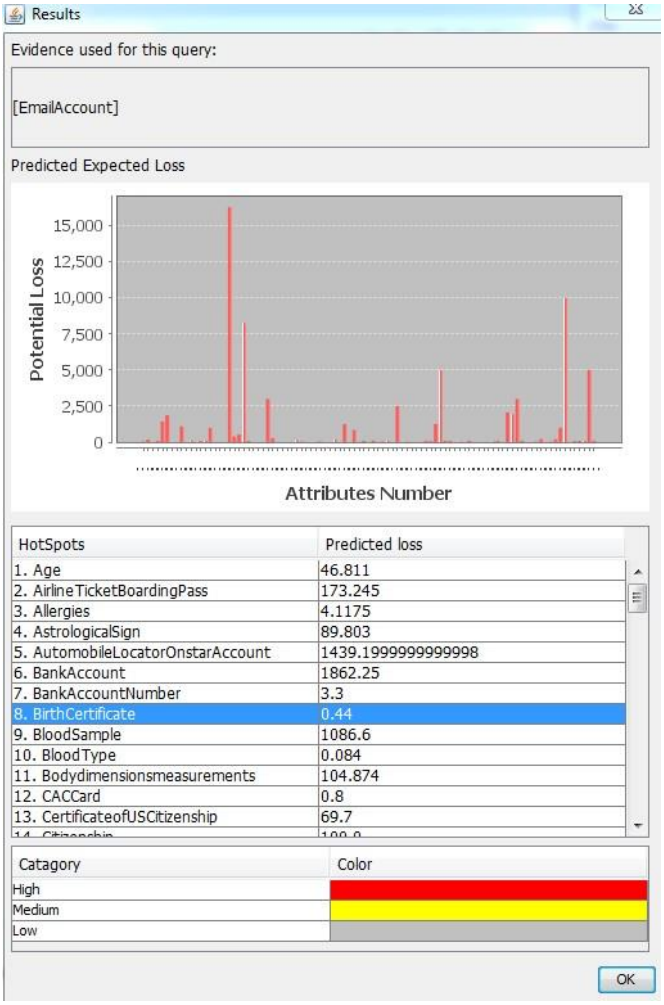

Figure 16: Results of Hotspots Query Represented in Bar Graph, with Multi-Dimension Specialization

Chapter 5: Experiments of Specialization on ITAP Data

Having demonstrated the functionality of specialization in the Identity Ecosystem based on Empirical data, this chapter focuses on the experiments based on data from real world provided by ITAP. Contents include introduction of ITAP data representation, learning general/specialized graphic models from ITAP data, and experiments of specialization based on graphic models obtained. It is important to note that a test data set is usually needed to estimate the accuracy of a graphic model. In this thesis, however, conducting such a test is meaningless due to the small amount of data available. The test will be future work when ITAP collects enough data.

5.1 ITAP DATA REPRESENTATION

The Identity Threat Assessment and Prediction (ITAP) project at the Center for Identity at the University of Texas is a project that represents and manipulate online fraud stories in a structured computational representation [7][8][26]. Each online story is represented as a "scenario". Figure 17 shows a typical scenario named "Albany Store Credit Card Fraud". It has information like inputs, which are attributes used by thieves, and outputs, which are outcomes or other attributes obtained (for the purpose of this thesis, outcomes are not used because they are just description of the results). It also has identity related information includes total loss (lossAmount) and information (ageGroupOfVictims, genderOfVictims, etc.) used for specialization purpose. One may notice that information of the scenario shown in Figure 17 is not complete, for example, educationLevelOfVictims and ProfessionOfVictims are unknown. In fact, Incompleteness is the nature of online stories. This property makes Bayesian Network

an ideal model for Identity Ecosystem, because of its ability of dealing with incomplete data.

```
- <scenario name="Albany Store Credit Card Fraud">
  - <inputs>
      <input name="Personally Identifiable Information (PII)" type="D"/>
      <input name="Store Credit Card(s) Opened" type="E"/>
    </inputs>
  - <outputs>
      <output name="Consumer Goods Purchased" type="E"/>
      <output name="Credit Card Charged" type="E"/>
    </outputs>
  - <identity>
      <criminalActivities>Outsider</criminalActivities>
      <ageGroupOfVictims>Adult</ageGroupOfVictims>
      <genderOfVictims>B</genderOfVictims>
      <citizenshipOfVictims>North America</citizenshipOfVictims>
      <educationLevelOfVictims>Unknown</educationLevelOfVictims>
      <annualIncomeOfVictims>10-20k</annualIncomeOfVictims>
      <professionOfVictims>Unknown</professionOfVictims>
      <organizationNameAndTicker>Various Stores (not named specifically in article)</organizationNameAndTicker>
      <locationOfEvent>Albany, NY</locationOfEvent>
      <dateWhenEventOccurred>Sep 2013</dateWhenEventOccurred>
      <dateOfArticleOrAnnouncement>07/02/2014</dateOfArticleOrAnnouncement>
      <lossIncurred>Financial</lossIncurred>
      <lossAmount>11029</lossAmount>
      <reputation>Low</reputation>
      <emotionalDistress>Low</emotionalDistress>
      <reported>Y</reported>
      <counterMeasuresTaken>Unknown</counterMeasuresTaken>
    </identity>
</scenario>
```

Figure 17: A scenario example

5.2 LEARNING GRAPHIC MODEL FROM ITAP DATA

As discussed earlier in section 3.5.2.2, there are three groups of parameters in the Identity Ecosystem graphic model, namely, conditional probability table (CPT), attribute values, and attribute priors (intrinsic probability that an attribute is exposed). Learning attribute priors from ITAP data is challenging. An obvious way is to count the ratio of number of scenarios one attribute appears as an input to the total number of scenarios. However, this is problematic because if an attribute is exposed it does not mean that attribute must appear in every scenario. In fact, in most cases an attribute is not in the input set of a scenario not because it is unknown by thieves, but that it is not needed in

the particular case. Here empirical data is utilized for the attributes. Remember that empirical attribute priors are needed for both the general model and each specialized model.

For the general model, both CPT and attribute values can be learned from ITAP data. Learning attribute values is straightforward. For a scenario s_i where "loss amount" is known, assign a loss to each of the attributes a_j appearing in inputs set or outputs set of s_i:

$$\text{Loss}_{s_i a_j} = \text{Loss}_{s_i} / N_{a s_i} \tag{12}$$

where Loss_{s_i} is the loss amount in s_i, N_a is the total number of attributes (or PII) appear in inputs set and outputs set in s_i. Thus the loss of attribute a_j can be calculated as:

$$\text{Loss}_{a_j} = \frac{\sum_{s_i \in S_{a_j}} \text{Loss}_{s_i a_j}}{N_{s a_j}} \tag{13}$$

where S_{a_j} is scenario set that include a_j as one of its inputs or outputs, $N_{s a_j}$ is the number of scenarios in set S_{a_j}. Here attributes in both inputs set and outputs set are considered, because attributes in both sets may affect the final loss amount.

Learning CPT is tricky. Ideal case is to learn conditional probability of each possible combination of parent attributes separately [9][10][27]. However, this would need massive amount of training data, which is not practical at the current stage. An alternative way is to assume that each parent attribute is independent from each other, and, therefore, conditional probability of attribute a_j for each single parent attribute $\text{p}_{a_j k}$ is calculated as shown in Equation 14:

$$\text{p}\left(a_j \middle| \text{p}_{a_j k}\right) = \frac{N_{a_j}}{N_{\text{p}_{a_j k}}} \tag{14}$$

where $N_{\text{p}_{a_j k}}$ is the total number of scenarios in set $\text{S}_{\text{p}_{a_j k}}$ (a set of all scenarios that contain $\text{p}_{a_j k}$ as one of their inputs). N_{a_j} is the number of scenarios in $\text{S}_{\text{p}_{a_j k}}$ such

that a_j is one of their outputs. The conditional probability of attribute a_j for a certain set of parent attributes $\mathrm{p}_{a_j} = \{\mathrm{p}_{a_j k}\}$ can be calculated as [6]:

$$\mathrm{P}\left(a_j \middle| \mathrm{p}_{a_j}\right) = 1 - \prod_{\mathrm{p}_{a_j k} \in \mathrm{p}_{a_j}} (1 - \mathrm{p}\left(a_j \middle| \mathrm{p}_{a_j k}\right)) \quad (15)$$

For specialized models, partial specialization technologies proposed in section 3.5.2 are applied, including partial specialization on both parameters and attributes. For parameters, attribute priors and attribute values are specialized and use the same CPT as in the general mode. As explained in section 3.5.2, this is reasonable because specializing CPT needs much larger amount of training data than attribute values, but the amount of data available is limited. Also, CPT is less likely to be different than attribute value, since thieves usually use similar methods for different groups of people, but it is people themselves that make their attribute values different. Specialization of attribute priors is straightforward: simply assign priors for each specialized model based on prior knowledge. Specialization related information in ITAP data can be used to group scenarios, each group of scenarios then can be used to populate their corresponding specialized model for attribute values using Equation 12 and Equation 13. For multi-dimension specialization, average attribute values and attribute priors of all related single-dimension specializations are used.

In this experiment, partial specialization of attributes is applied to attribute priors, but not to attribute values. It is much easier to determine if an attribute prior is affected by a specialization criterion than that of an attribute value, simply because the former is based on prior knowledge. In addition, all attribute values are specialized so that one can figure out how different an attribute value will be in each specialized model. This information may be useful in specializing models in the future.

5.3 RESULTS

5.3.1 Reduced Specialization Criteria and Attributes List

In this experiment, only 5 of the 7 specialization criteria proposed in Table 1 are used. “Profession” and “Citizenship” are discarded because specialization information from ITAP cannot be obtained. In addition, currently only 791 stories are manually constructed as scenarios by ITAP project, with only part of them are related to individuals (the others are related to device and companies, which is not the focus of this thesis). Thus, the number of attributes involved is smaller than the original attribute list in the Identity Ecosystem. A reduced attributes list with a total number of 68 is used in this experiment. More than half of the attributes in the reduced attributes list are directly from the original list. The rest may be new attributes discussed in the 791 scenarios used here, or merged attributes from attributes in the original list.

5.3.2 Appling Specialization

As mentioned in section 5.2, “partial specialization on attributes” is applied to attribute priors. Table 3 shows statistics of specialized attributes for each specialization criterion applied. “Age” appears to be the most efficient criterion.

Table 3: Statistics of specialized attributes in specialization criteria used

Criterion	Age	Gender	Education	Income	Citizenship
No. of Attributes specialized	59	21	45	46	17
Percentage (%)	86.7	30.1	66.2	67.7	25

Table 4: Part of learned attribute values for specialization criterion “Age”

	Baby	Child	Teen	Adult	Senior	Deceased
CPTCode	0	0	0	0	0	0
IDCard	0	0	0	43965.45	0	0
IDNumber	0	0	0	5000	0	0
VIN	0	0	0	0	0	0
ZIPCode	0	0	0	0	0	0
address	0	0	0	81375	3000	9162.5
age	0	0	0	0	66000	9162.5
bankAccount	0	0	0	2007770.2	1478500	0
barLicense	0	0	0	0	0	0
billRecords	0	0	0	14583.333	0	0
biographicData	0	121400	0	421152.2	82553.5	101548.8
biometricData	0	0	0	0	0	0
birthCertificate	0	0	0	2473	0	0
carPurchasingInformation	0	0	0	750	0	0
check	0	0	0	190464.11	26525.83	0
citizenship	0	0	0	0	0	0
contacts	0	0	0	0	0	0
courseSchedule	0	0	0	0	0	0
creditCard	0	121400	0	41517.855	25640.08	0
creditInformation	0	0	0	0	0	0
date	0	0	0	0	0	9162.5
dateOfBirth	0	0	0	127561.9	0	9162.5
deathCertificate	0	0	0	6000	0	0
debitCard	0	0	0	664238.9	135000	0
driverLicense	0	0	0	291788.12	162107	0
electronicBenefitsTransferAccount	0	0	0	3333.3333	0	0
email	0	0	0	0	0	0
employeeinFormation	0	0	0	82869.05	0	0
evictioninFormation	0	0	0	0	0	0

Table 4 shows learned attribute values for part of the attributes with specialization criterion “Age”. “0” value means no information is available from the ITAP data. Table 4 indicates two points: 1. the amount of training data currently available is far from large enough to populate all specialized attribute values. 2. The data is not evenly distributed to

different specialization groups of “Age”. The first point means more training data are needed in order to conduct a complete experiment. The second point is disappointing because it means even more data than we expected are required for minor specialization groups such as “Baby” and “Teen”. However, it is nice to see that there is more data than expected for major specialization groups like “Adult” and “Senior”, which makes it possible to conduct experiments on those major specialization groups at the current stage.

Table 5: Percentage of successfully populated attribute values in specialized models

Age	Baby	Child	Teen	Adult	Senior	Deceased
Percentage(%)	0	2.9	0	48.5	17.6	13.2
Annual Income	<10K	10-20K	20-50k	50-100k	100-500k	>500k
Percentage(%)	5.9	4.4	10.3	5.9	0	16.2
Citizenship	North America	South America	Europe	Asia	Africa	Australia
Percentage(%)	64.7	0	2.9	11.8	1.5	0
Education	Uneducated	Primary School	Secondary School	High School	College	Graduate School
Percentage(%)	2.9	0	2.9	4.4	30.9	4.4
Gender	Male	Female				
Percentage(%)	58.8	22.1				

The situation encountered by criterion “Age” is also true for other specialization criteria applied. Table 5 shows statistical results of attribute values successfully populated by ITAP data. Most specialized models are far from fully populated. However, there are still some specialized models that are “usable”, they are “Adult” in “Age”, “North America” in Citizenship, and “Male” in “Gender”. Experiments on queries of specialized models in section 5.3.4 will be based on these “usable” specialized models. For comparison, the percentage of attribute values successfully populated for the general model is 64.7%.

The CPT used for specialized models, as explained in section 5.2, is the same one that is used for the general model. With the 791 scenarios used, there are 295 edges

obtained. This means that our graphic model is a sparse one. With 68 nodes, there are $67 \times 68 = 4556$ possible edges.

5.3.3 Model Preprocessing

There are several scenarios that have super high loss (outliers), which leads to super high attribute values (tens of millions) for three of the attributes: intellectual property information , SSN and user account. Query results based on such data will be overwhelmed by these attributes with super high value. Here the maximum loss per scenario is limited to be 1 million. If a loss exceeds this value, simply use 1 million as its loss.

In addition, the attribute "credit card" has a number of parents as high as 26, which means it needs $2^{26} = 67108864$ conditional probabilities. The current implementation of the Identity Ecosystem calculated all conditional probabilities and stored them before queries. Thus the large number of conditional probabilities needed by "credit card" may suspend the process. Here 8 less likely links pointing to "credit card" are deleted before queries, for example, link from "paper work" to "credit card". In future experiments, large number of parents for attributes is almost inevitable. A solution could be, instead of calculating all conditional probabilities and storing them, only calculate a conditional probability when it is needed for a query. This approach would require modification to the implementation of the Identity Ecosystem.

5.3.4 Experiments on Queries

This section compares query results from the general models and specialized models. To save space, only PII Exposure Query is conducted here. Comparison of two

other queries will be similar. Specialized models are chosen based on the statistics shown in Table 5: simply choose the models that have high percentages of populated attribute values. Figure 18 shows the query results from the general model with evidence “email”. From the results we see SSN and tax information is at the highest risk. The general mode, although may not be accurate due to the small amount of training data, shows a relatively complete relationship among different attributes.

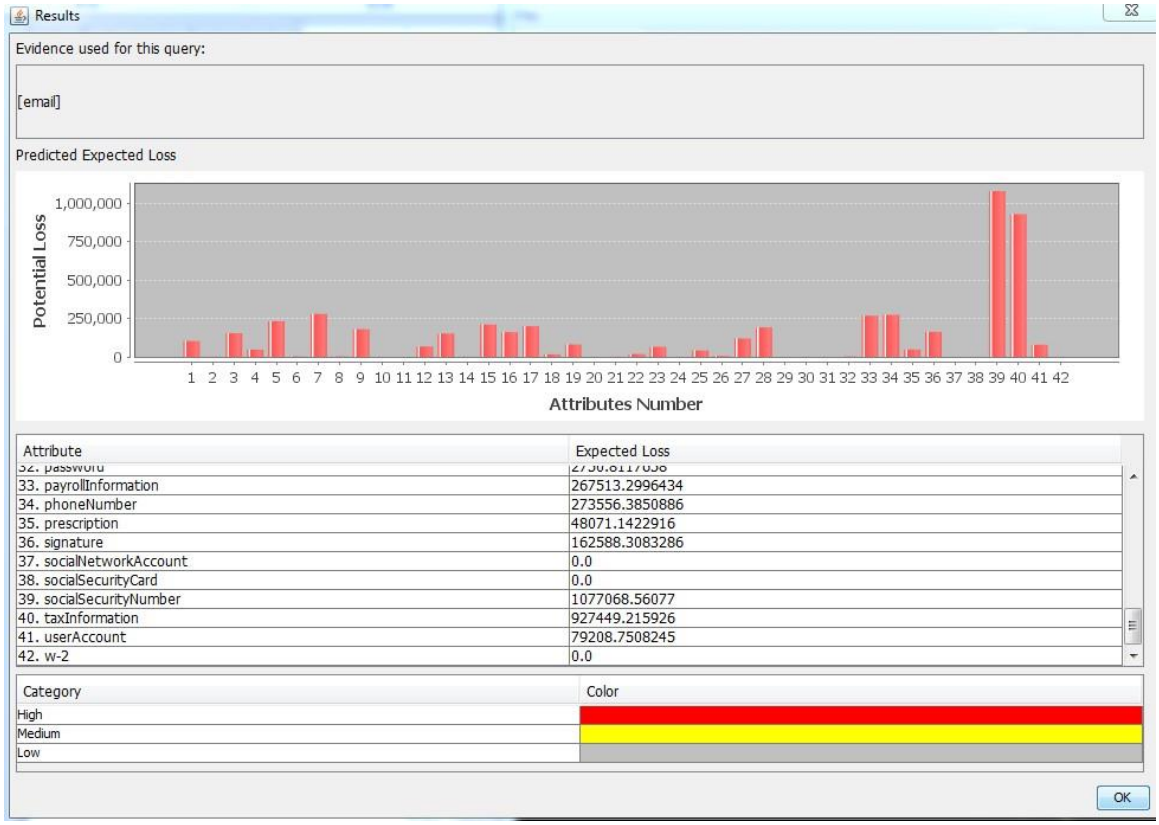

Figure 18: Results of PII Exposure Query from the general model, with evidence “email”

Figure 19 shows the query results from a one-dimension specialized model, with specialization “Age: Adult” and evidence “email”. It has a similar pattern as those results from the general model, because they share the same CPT. However, its loss values are

much smaller. This implies that attribute values for this specialized model related to attribute “email” are not appropriately learned, probably because of lacking specialization information for “email” related attributes.

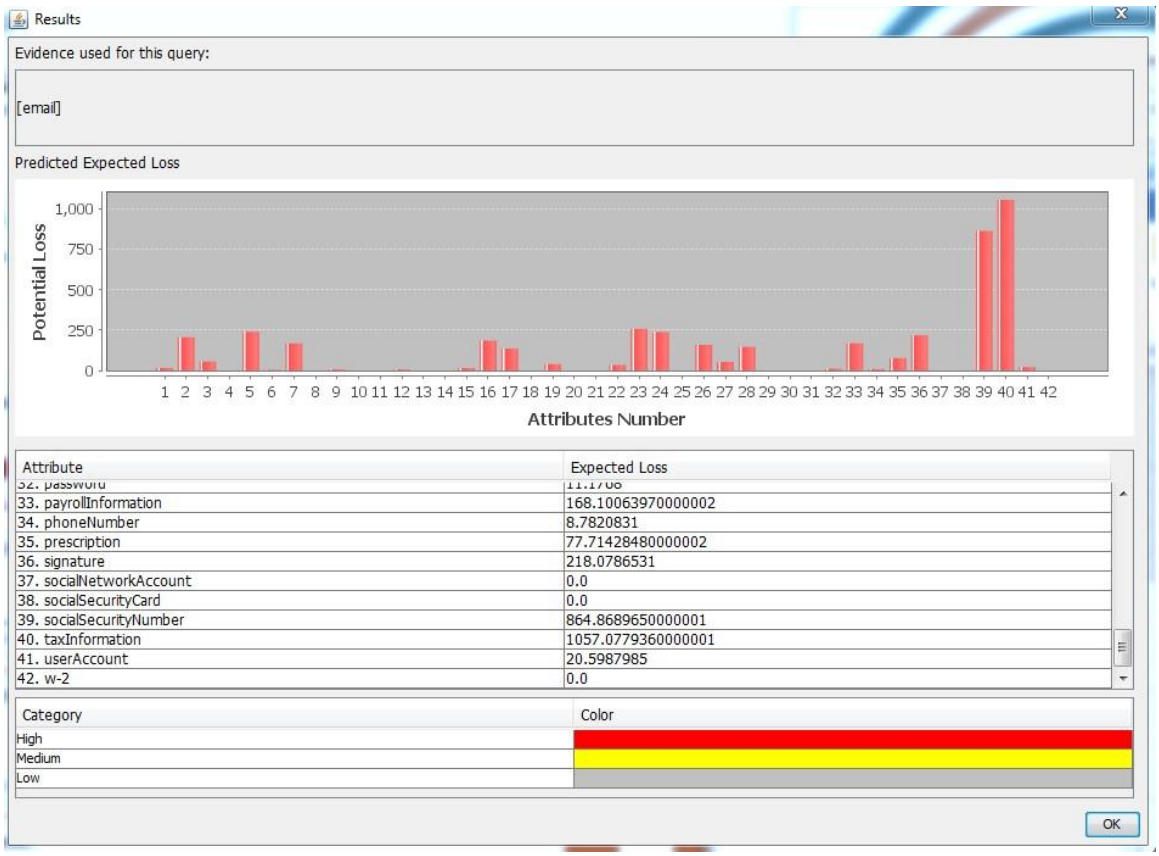

Figure 19: Results of PII Exposure Query from one-dimension specialized mode with specialization criterion “Age” and group name “Adult”, with evidence “email”

Figure 20 shows the query results from a one-dimension specialized model, with specialization “Gender: Male” and evidence “email”. Both of its distribution pattern and loss values are similar to the general model. This is expected because: 1. They share the

same CPT, 2. There are only two groups “Male” and “Female”, in which group “Male” utilizes the majority of training data (as shown in Table 5).

Figure 20: Results of PII Exposure Query from one-dimension specialized mode with specialization criterion “Gender” and group name “Male”, with evidence “email”

Figure 21 shows results of PII Exposure Query from Multi-dimension specialized mode with specialization “Age: Adult”, “Gender: Male” and “Citizenship: North America”, with evidence “email”. Currently, the multi-dimension is simply average of the one-dimension specialized models involved. Due to the small amount of training data, the results, as expected, are closer to those of the general model than those of corresponding one-dimension specialized models.

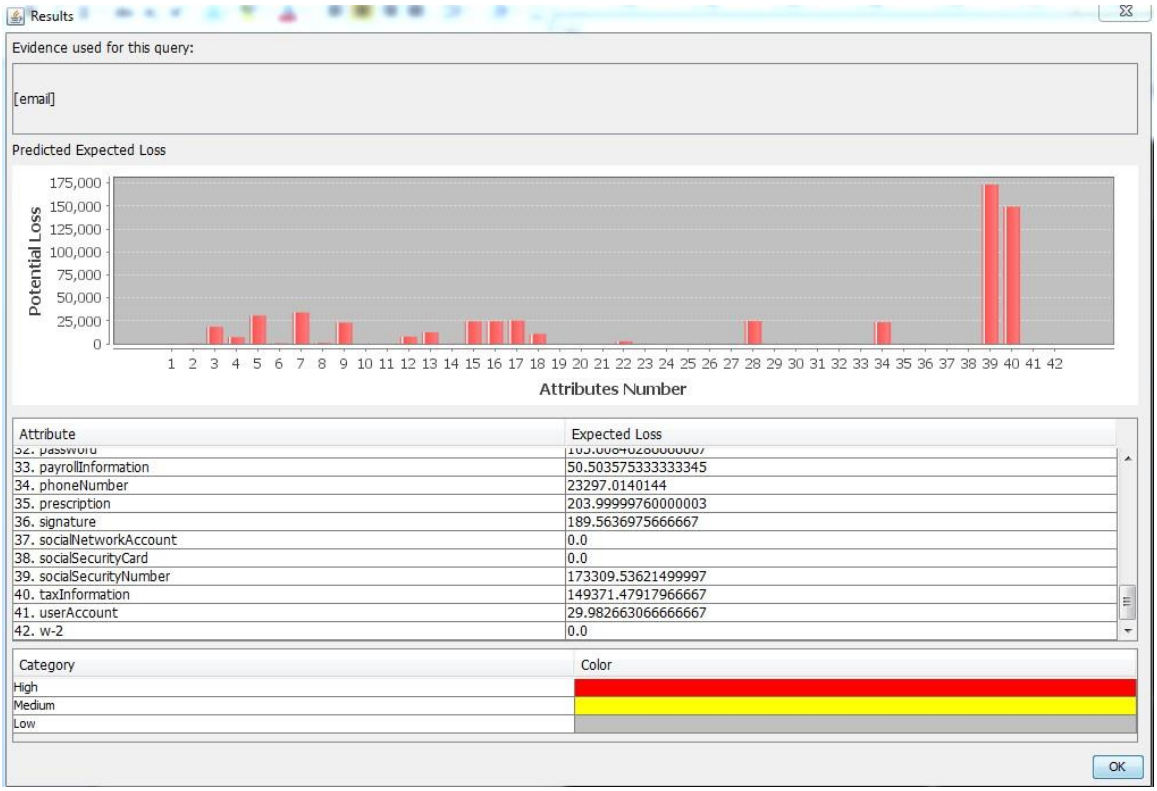

Figure 21: Results of PII Exposure Query from Multi-dimension specialized mode with specialization “Age: Adult”, “Gender: Male” and “Citizenship: North America”, with evidence “email”

In conclusion, the general model is “reasonably” populated with the 791 scenarios. That is to say, relationship among attributes is reasonably discovered but model parameters may not be accurate enough. Specialized models, however, are generally far from fully populated. Much more training data is needed to fully exam the effect of specialization. As the ITAP project continues collecting data from online stories, this task can be fulfilled in the near future.

Chapter 6: Conclusion and Discussion

In this thesis, specialization of graphic model in the Identity Ecosystem is investigated. Three research problems related to specialization are carefully studied in chapter 3: 1. how to choose specialization criteria and their boundaries, 2. how to achieve multi-dimension specialization based on one-dimension specialization, and 3. how to overcome the tradeoff between the desire of high specialization degree and limited amount of data. Theories and technologies developed in chapter 3 are then applied to experiments based on empirical data (chapter 4) and ITAP data (chapter 5). The difficulty in this research is the insufficient ITAP data currently available. Thus, experimental results on ITAP data in chapter 5 are not accurate. However, specialized graphic models are supposed to provide more accurate prediction in their corresponding scope given enough training data. This research develops theories and technologies of specialization in a general way: discuss different cases with various amounts of training data. For example, partial specialization technologies developed in chapter 3 are quite flexible when applied. CPT, attribute priors and attribute values can be chosen to be specialized or not, to fit each special case with different amounts of training data. Different ways of applying multi-dimension specialization are also discussed to fit different situations. Finally, theories developed for determining boundaries of specialization criteria are expected to be applied for future research when more training data is available. As the ITAP project continues collecting data, theories and technologies developed in this research can be customized along the way to fit the needs. In summary, works in this thesis are expected to be a guideline of designing specialization in future works.

Chapter 7: Future Work

There are still many research topics in specialization worth further investigation. First of all, it may be helpful to apply classifiers with nonlinear boundaries to multi-dimension specialization when more data is available. The grid like boundaries currently used is quite weak. They may not be able to capture significant features of the distribution of data points in the specialization space. In addition, using different kinds of boundaries also gives freedom to choose the number of specialized groups. For example, with nonlinear boundaries, or even linear boundaries, one may only divide data points to two specialized groups in three-dimension specialization with criteria: "age", "income" and "gender", as long as it makes sense. In this case, any combination of the three criteria will fall into one of the two groups. With grid like boundaries, however, one needs to divide data points into at least 8 groups, when each criterion has 2 groups. This disadvantage significantly limits the usage of specialization with limited amount of data.

Another problem worth further study is how to represent correlation of two specialization criteria in mathematic form. In section 3.5.1, eliminating correlations among specialization criteria is proposed as a solution for the trade-off. However, correlations are introduced by reasoning, rather than a measurable value. Representing correlations of specialization criteria in appropriate math form help one optimize the choice of specialization criteria in multi-dimension specialization.

In addition, the technology "partial specialization on attribute values" may be applied in future work when more data are available. Of course, there is no need to partial specialize attribute values if arbitrary amount of training data is available. The situation discussed here is that the amount of data is not enough to populate attribute values of

each specialized models, but is large enough to figure out a group of attributes whose values are less necessary to be specialized than the rest.

Finally, the structure of ITAP data should be modified for future specialization design. Currently most specialization related information in scenarios is specialized group names. Nevertheless, optimizing of specialization boundaries requires detailed information. For example, one may need the exact age instead of established age groups like “Child”, “Adult”, etc. so that one can regroup people in an optimized way.

Bibliography

[1] E. Schonfeld. (2011, December) Android phones pass 700,000 activations per day, approaching 250 million total. [Online]. Available: http://techcrunch.com/2011/12/22/android-700000/

[2] Duncan Hodges, Sadie Creese and Michael Goldsmith, "A Model for Identity in the Cyber and Natural Universes" 2012 European Intelligence and Security Informatics Conference.

[3] Federal Trade Commission. 2006 Identity Theft Survey Report. General Books LLC, 2007.

[4] Javelin Strategy and Research. 2011 identity fraud survey report. JSR, 2011.

[5] Graeme R. Newman and Megan M. McNally. Identity Theft Literature Review. National Institute of Justice, 2005.

[6] Liang Zhu, Shayani Deb, Muhammad Zubair Malik and Suzanne Barber, "Predicting and Explaining Identity Risk, Exposure and Cost using Ecosystem of Identity Attributes", to be submitted.

[7] Yongpeng Yang, "Mining of Identity Theft Stories to Model and Assess Identity Threat Behaviors", Master's Thesis, University of Texas at Austin, 2014.

[8] Yongpeng Yang, Monisha Manoharan, and Suzanne Barber, "Modelling and Analysis of Identity Threat Behaviors through Text Mining of Identity Theft Stories", accepted for publication in IEEE Joint ISI 2014.

[9] David Heckerman, "A tutorial on learning with Bayesian networks," Technical report, Microsoft Research, 1996.

[10] Christopher Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.

[11] Bayesian Network Wiki: http://en.wikipedia.org/wiki/Bayesian_network

[12] Nevin Lianwen Zhang, and David Poole, "A simple Approach to Bayesian Netowrk Computations",http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.70.622&rep=rep1&type=pdf

[13] S. L. Lauritzen, and D. J. Spiegelhalter, "Local Computation with Probabilities on Graphical Structures and their Application to Expert System", Vol. 50, 157-224 (1988).

[14] Finn. Jensen, S. Lauritzen, and K. Olesen, "Bayesian updating in recursive graphical models by local computations", Computational Statisticals Quarterly, Vol. 4, 269-282 (1990).

[15] A. P. David, "Applications of a general propagation algorithm for probabilistic expert systems", Statistics and Computing, Vol. 2, 25-36 (1992).

[16] Treewidth Wiki: http://en.wikipedia.org/wiki/Treewidth

[17] B. J. Frey, and D. J. C. Mackay, "A revolution: Belief propagation in grpahs with cycles", Advances in Neural Information Processing System, Vol. 10. MIT Press (1998).

[18] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss, "Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms", IEEE Transactions on Information Theory, Vol 51, 2282-2312 (2005).

[19] T. Minka, "Expectation Propagation for Approximate Bayesian Inference", Proceedings of the Seventeenth Conference on University in Artificial Intelligence, 362-369 (2001).

[20] T. Minka, "A Family of Approximate Algorithms for Bayesian Inference", Ph.D. thesis, MIT (2001).

[21] Importance sampling Wiki: http://en.wikipedia.org/wiki/Importance_sampling

[22] Gibbs sampling Wiki: http://en.wikipedia.org/wiki/Gibbs_sampling

[23] Specialization Wiki: http://en.wikipedia.org/wiki/Specialization

[24] K. Fukunaga, "Introduction to Statistical Pattern Recognition", Second ed., Academic Press.

[25] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, "The Elements of Statistical Learning: Data mining, Inference, and Prediction", New York: Springer (2009).

[26] "The ITAP: Getting One Step Ahead", inner material in the Center for Identity, University of Texas at Austin.

[27] David Heckerman, Dan Geiger, David M. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," Machine Learning, 20, 197-243 (1995).